Manuscript in press at *Cortex*

Within-category representational stability through the lens of manipulable objects

Dongha Lee[1,2,3], Jorge Almeida[1,2]*

1. Proaction Laboratory, Faculty of Psychology and Educational Sciences, University of Coimbra, Portugal
2. CINEICC, Faculty of Psychology and Educational Sciences, University of Coimbra, Portugal
3. Korea Brain Research Institute, Daegu, Republic of Korea

* Corresponding Author:
Jorge Almeida
Faculty of Psychology and Education Sciences,
University of Coimbra,
Rua do Colégio Novo, 3001-802 Coimbra
Email: jorgealmeida@fpce.uc.pt

Abstract

Our ability to recognize an object amongst many exemplars is one of our most important features, and one that putatively distinguishes humans from non-human animals and potentially from (current) computational and artificial intelligence models. We can recognize objects consistently regardless of when we see them suggesting that we have stable representations across time and different contexts. Importantly, little is known about how humans can replicate within-category object representations across time. Here, we investigate neural stability of within-category object representations by computing the similarity between representational geometries of activity patterns for 80 images of tools obtained on different fMRI scanning days. We show that within-category representational stability is observable in regions that span lateral and ventral temporal cortex, inferior and superior parietal cortex, and premotor cortex – regions typically associated with tool processing and visuospatial processing. We then focus on what kinds of representations best explain the representational geometries within these regions. We test the similarity of these geometries with those coming from the different layers of a convolutional neural network, and those coming from perceived and veridical visual similarity models. We find that regions supporting within-category representational stability show stronger relationship with higher-level visual/semantic features, suggesting that neural replicability is derived from perceived and higher-level visual information. Within category representational stability may thus originate from long-range cross talk between category-specific regions (and in this case strongly within ventral and lateral temporal cortex), over more abstract, rather than veridical/lower-level, visual (sensorial) representations, and perhaps in the service of object-centered representations.

Keywords:
Representational stability; within-category tool representations; fMRI; CNN; perceived similarity

Introduction

Object recognition is a computational hard problem that humans solve, seemingly effortlessly, every day, and that requires stable representations at different levels of generalization (e.g., Baylis & Driver, 2001; Booth & Rolls, 1998; DiCarlo & Cox, 2007; Grill-Spector et al., 1999; Konen & Kastner, 2008; Pourtois, Schwartz, Spiridon, Martuzzi, & Vuilleumier, 2009; Rollenhagen & Olson, 2000). We can certainly recognize a hammer as a manmade object, as a graspable object, or as a tool (i.e., at the superordinate/category level), as well as a hammer (i.e., basic/type level) or even as my hammer (versus my neighbor's hammer; i.e., the subordinate/exemplar level), and we do so consistently across time. This representational stability provides the cognitive system with sufficiently detailed information for adequate and consistent discriminability of and disambiguation between different objects, irrespective of the current context. Importantly, different levels of categorization (e.g., superordinate versus basic level) will need different grains of information to be replicable at different times in order to accurately categorize and recognize a stimulus – for instance, the degree of informational specificity required to categorize an object as my hammer may be orders of magnitude higher than that necessary for categorizing a hammer as a tool. Here we focus on how stable are object-specific within-category neural representations across time by looking at how similar neural patterns for a set of individual tools are in two functional magnetic resonance imaging (fMRI) sessions separated by at least a week.

Object identification and object categorization are perfect examples of processes that need different levels of generalization. In object categorization, when particular criteria at the superordinate level are satisfied objects are recognized as members of a single category (e.g., criteria: inanimate, graspable, manmade; categorization output: tool) (Serre, Oliva, & Poggio, 2007; Warrington & McCarthy, 1987). In object identification, however, objects are recognized specifically when considering a set of object-specific features at the basic level (e.g., criteria: inanimate, graspable, manmade; features: it cuts, it has a blade, it has a handle, it is serrated, it is used in the kitchen; identification output: knife) (Gerlach & Marques, 2014; Riesenhuber & Poggio, 2002; Tyler & Moss, 2001). Although these are equally important and demanding processes, we go about our world mostly by identifying objects rather categorizing them. Surprisingly, the majority of the studies hitherto have focused on object representations across a fixed set of categories in lieu of exploring within-category representations (Charest, Kievit, Schmitz, Deca, & Kriegeskorte, 2014; Cichy, Khosla, Pantazis, Torralba, & Oliva, 2016; Kriegeskorte, Mur, & Bandettini, 2008; Kriegeskorte, Mur, Ruff, et al., 2008; Mur et al., 2013). In spite of variance in structural dimensions within a category, object representations have been explored (Cree & McRae, 2003; Marques, Raposo, & Almeida, 2013). For instance, Charest and colleagues (Charest et al., 2014) showed that representational geometry for objects from different categories (e.g., bodies, faces, places, objects) remained consistent in early visual

cortex and inferior temporal cortex across different scanning days.

One unaddressed question then is how stable are within-category object-specific neural representations. When identifying an object, we need to gather fine-grain, perhaps multimodal and conceptual information in order to choose the target object type amongst other candidate object types and reach an identification decision (e.g., choose a hammer from a set of possible objects such as an axe, a knife, a screwdriver, etc.). In fact, it is perhaps possible that visual information, although extremely informative for some object categorization decisions (e.g., Almeida et al., 2014; Cree & McRae, 2003; Marques et al., 2013; Sakuraba, Sakai, Yamanaka, Yokosawa, & Hirayama, 2012), is not sufficient for object type identification, and that high-level sensory (visual) and non-sensory information is needed for (within-category) object identification. Moreover, it is also probable that certain aspects of conceptual representation – aspects that are constitutive of the concept – may be stable across different instantiations, and may be grounded by other object-related information that are triggered by context and the situational aspects at play at the specific time the concept is instantiated (e.g., when KNIFE is presented in the context of us wanting to spread some butter, cut a juicy piece of meat, or sharpen a wooden stick for toasting marshmallows different instantiations may be achieved) (e.g., Mahon & Caramazza, 2008). Thus, looking at conceptual stability (through the lens of representational stability) across different instantiations of an object (e.g., the same object being recognized at different times, under different contexts and situational pressures) may give us leverage over the complex questions of object representation.

In the current study, we examine what factors contribute to neural stability of within-category object-specific representations using fMRI and Representational Similarity Analysis (RSA) (e.g.,Kriegeskorte, Mur, Ruff, et al., 2008; Lee, Mahon, & Almeida, 2019). Specifically, we focus on how replicable are tool representations across the brain by inspecting how similar multivariate response patterns are for 80 individual tool items across two independent fMRI scanning sessions separate by at least a week. We will then focus on whether regions that show high within-category representational stability are better explained by different levels of visual features (i.e., lower to higher visual features) by comparing neural similarity within these regions with object-specific similarity at different layers of a convolutional neural network (CNN; Krizhevsky, Sutskever, & Hinton, 2012). Finally, we will test whether neural similarity in these regions is dictated by perceived or veridical (i.e., pixel-to-pixel) visual similarity between the tools used. We predict that regions within temporal and parietal and premotor associative cortex, typically involved in high-level processing of tools (e.g., Almeida, Fintzi, & Mahon, 2013; Chao, Haxby, & Martin, 1999; Chao & Martin, 2000; Garcea, Kristensen, Almeida, & Mahon, 2016; Kristensen, Garcea, Mahon, & Almeida, 2016; Lee et al., 2019; Mahon, Kumar, & Almeida, 2013; Mahon et al., 2007; Noppeney, Price, Penny, & Friston, 2006) will show high within-category (tool) object-specific representational stability across

fMRI sessions, as some of these regions represent amodal and stable properties of objects. Moreover, we predict that the representational geometries in these regions are better explained by later layers of the CNN and by perceived similarity between the tools, suggesting that within-category replicability is dependent on higher-level abstract visual and conceptual representations.

Results

Neural responses to tools were measured using whole brain fMRI in two different sessions acquired in different days. All participants completed 5 runs per session. Each participant was presented with 80 tool images, with different exemplars presented for odd and even runs. We extracted activity patterns from the neural responses in predefined regions-of-interest (ROIs). We used the human Brainnetome Atlas composed of 246 ROIs within the two hemispheres to obtain more-fine grained brain regions, and used RSA to compute session-specific representational similarity matrices (RDMs) and then extracted the representational structure of tool representations.

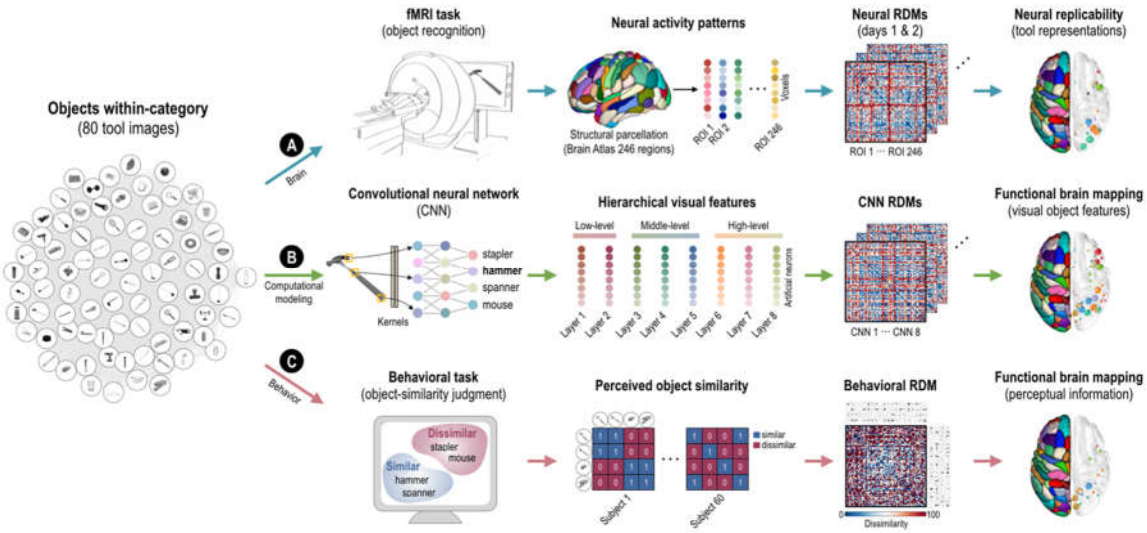

Figure 1. Overview of experimental procedures and data processing pipeline. The experiment was composed of three different representations: (i) neural activity, (ii) hierarchical visual feature, and (iii) perceptual information. **A)** Twenty participants performed an object recognition task. Neural activity patterns for 80 tools were obtained from each brain region and neural RDMs were calculated based on the dissimilarity of all pairs of activity patterns; **B)** We used a convolutional neural network (CNN) to extract visual features of objects in low-, middle-, and high-level layers. CNN RDMs were calculated by the correlation distance between visual features; **C)** Sixty participants participated in an object-similarity judgment task in order to obtain perceptual representations Each participant's sorting scores (i.e., the value is 1 if two presented tools are in the same pile, and zero if they are not) were averaged across participants to define perceived object similarity. All RDMs were compared each other by calculating the correlation between RDMs.

Neural stability of within-category object-specific representations

To test how replicable tool representations are across the brain, we averaged neural RDMs across subjects (N=20) for each cortical region and session. The group-averaged RDMs were compared between the sessions using Pearson correlation. Figure 2 presents cortical brain regions showing high within-category representational stability (Bonferroni corrected at $p <$ 0,05). All statistical inferences of neural stability in these regions are summarized in Table 1. Importantly, regions (out of the 246 ROIs used) that show high stability between the two

sessions include regions within the fusiform gyrus on the left, lateral temporal cortex bilaterally, angular gyrus and superior parietal lobule on the right and ventral premotor bilaterally – all regions typically involved in the processing of tools (left fusiform gyrus: Chao et al. (1999), Lee et al. (2019), Mahon et al. (2007); lateral temporal cortex: Almeida et al. (2013), Chao et al. (1999), Mahon et al. (2007); right superior parietal: Almeida et al. (2013), Garcea et al. (2016), Kristensen et al. (2016), Mahon et al. (2013); ventral premotor cortex: Binkofski and Buccino (2006)), and action and spatial processing (right angular gyrus: Farrer et al. (2008)). Moreover, early visual cortical (EVC) regions also show high stability.

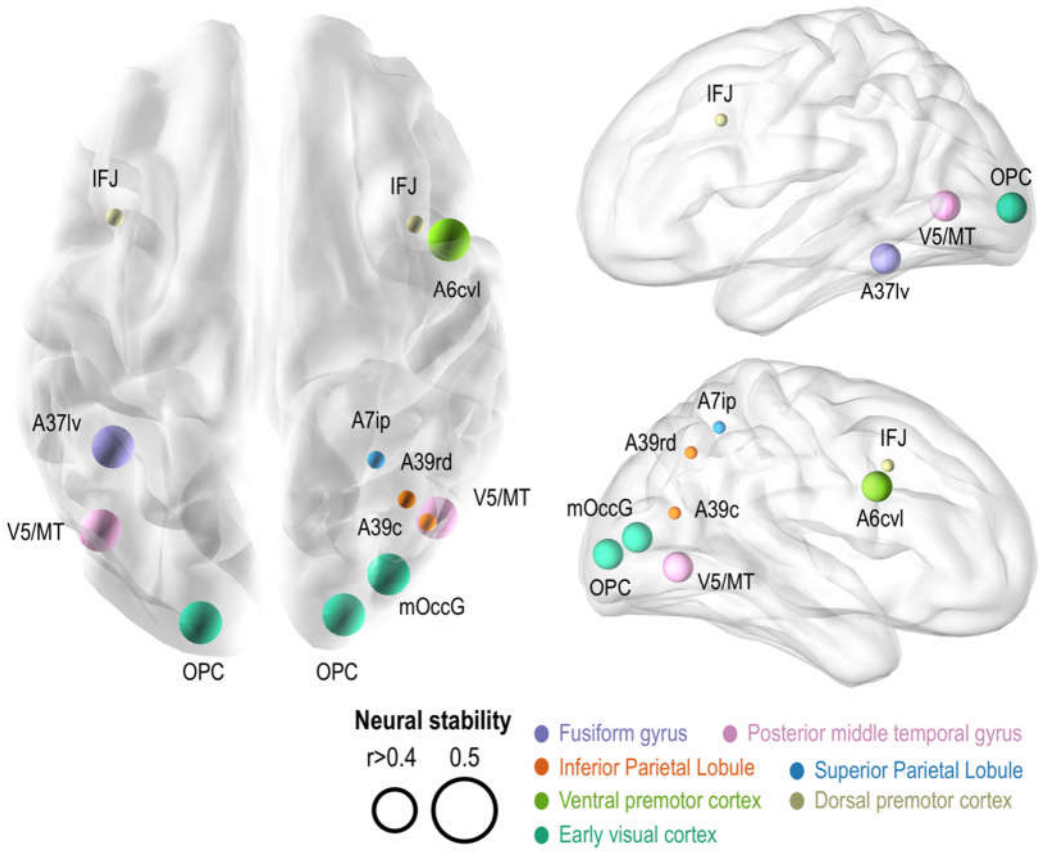

Figure 2. Distribution of brain regions with high neural stability of within-category object representations. Brain regions within temporal, frontal, visual, and parietal associative cortex show high neural stability across fMRI sessions. Significant brain nodes are displayed (Bonferroni-corrected $p < 0.05$).

Table 1. Brain regions showing high within-category representational stability (the top 5% threshold of correlation coefficients, total 12 regions remained). L: left, R: right.

Lobe	ROIs	Anatomical and modified Cytoarchitectonic descriptions from Brainnetome Atlas	Representational stability (Pearson r)	MNI (X,Y,Z)
Occipital Lobe	Early visual cortex	L.OPC (occipital polar cortex)	0.5857	−18, −99, 2
Temporal lobe	Fusiform gyrus	L.A37lv (lateroventral area 37)	0.5651	−42, −51, −17

Frontal lobe	Ventral premotor cortex	R.A6cvl (caudal ventrolateral area 6)	0.5553	51, 7, 30
Occipital Lobe	Early visual cortex	R.OPC	0.5531	22, −97, 4
Temporal lobe	Posterior middle temporal gyrus	R.V5/MT+	0.5368	48, −70, −1
Temporal lobe	Posterior middle temporal gyrus	L.V5/MT+ (area V5/MT+)	0.5309	−46, −74, 3
Occipital Lobe	Early visual cortex	R.mOccG (middle occipital gyrus)	0.5226	34, −86, 11
Parietal lobe	Superior parietal lobule	R.A7ip (intraparietal area 7, hIP3)	0.4983	31, −54, 53
Parietal lobe	Angular gyrus	R.A39c (caudal area 39, PGp)	0.4977	45, −71, 20
Frontal lobe	Dorsal premotor cortex	R.IFJ	0.4863	42, 11, 39
Parietal lobe	Angular gyrus	R.A39rd (rostrodorsal area 39, Hip3)	0.4759	39, −65, 44
Frontal lobe	Dorsal premotor cortex	L.IFJ (inferior frontal junction)	0.4609	−42, 13, 36

Different involvement of visual features in within-category high stability representational areas

We have shown which brain areas present high neural stability of within-category (tool) representations. We then asked what types of (visual) features subserve representations within regions that show within-category high representational stability. To do so, we tested the similarity between neural RDMs within each high-stability area and the RDMs derived from the different CNN layers (layers 1 through 8). As shown in Figure 3, all areas but the early visual cortical ones are show similarity with the representations in the higher-level layers (CNN layers 6 to 8), whereas visual features within the mid-level layers (CNN layers 3 and 4) showed greater similarity with representations in EVC. All statistical inferences of the similarity between neural and CNN RDMs in within-category high representational areas (Bonferroni-corrected $p < 0.05$) are summarized in Table 2.

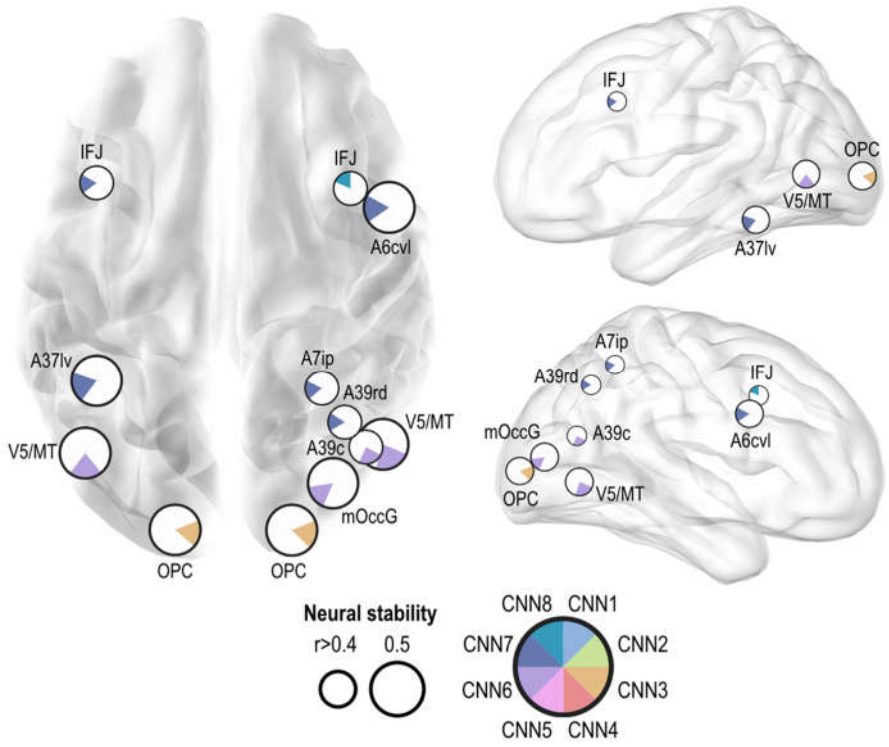

Figure 3. Distributions of hierarchical visual features in within-category high representational stability areas. The most significant layer is displayed using the color-coded pie chart. The size of the slice relates to the strength of correlation (Bonferroni-corrected $p < 0.05$).

Involvement of Perceived and Veridical visual similarity in within-category high representational stability areas

We further tested whether neural similarity in these regions is driven by perceived or veridical (pixelwise) visual similarity between the target tools. Importantly, and as can be seen in Figure 4, in EVC representational similarity was dependent on pixelwise veridical, rather than perceived, visual similarity between the presented tools, whereas in the remaining high-stability regions perceived visual similarity was predominant. All statistical inferences of the involvement of perceived and veridical visual similarity in within-category high representational stability areas (Bonferroni-corrected $p < 0.05$) are summarized in Table 2.

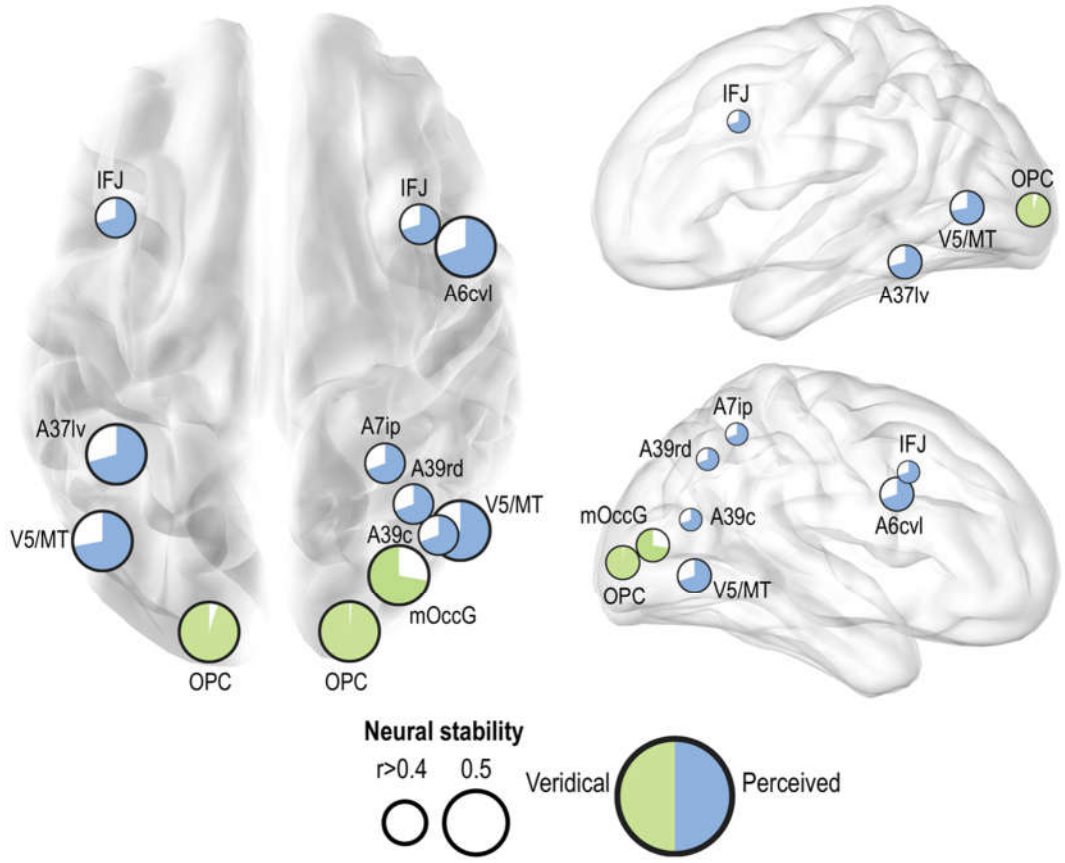

Figure 4. Distributions of perceived and veridical visual features in high within-category representational areas. The most significant feature is displayed using the color-coded pie chart. The size of the slice relates to the strength of correlation (Bonferroni-corrected $p < 0.05$).

Discussion

We set out to investigate stability in within-category object-specific representations using neural, computational, and behavioral models. This is a central issue for system's neuroscience (Kriegeskorte, Mur, & Bandettini, 2008) as it enables our understanding of how we go about our world recognizing and identifying objects. We first showed that within-category tool representations are stable across different timepoints. This stability is focused on regions typically involved in the processing of tools, such as ventral and lateral temporal cortex, superior parietal cortex, and ventral and dorsal premotor regions (Almeida et al., 2013; Chao et al., 1999; Chao & Martin, 2000), in regions dedicated to semantic processing of concrete objects and visuospatial and attentional processing of objects such as the right angular gyrus and right parietal cortex (Chen, Weidner, Vossel, Weiss, & Fink, 2012; Peelle, Troiani, & Grossman, 2009; Sabsevitz, Medler, Seidenberg, & Binder, 2005; Seghier, 2013), and the EVC. Importantly, this stability seems to be stronger in ventral stream regions – regions that are typical of tool processing such as the medial fusiform gyrus and the posterior middle temporal gyrus. Importantly, the representations within these areas seem to be dictated by abstract visual representations, as demonstrated by the fact that RDMs from these areas correlate with RDMs from later layers of a CNN and with RDMs extracted from perceived visual similarity measures rather than veridical visual similarity. This was true of all high stability regions except EVC, where representations seemed to be dictated by veridical visual similarity and earlier layers of the CNN.

Our data are, in part, in line with the reports on cross-category representation stability that show that regions within the inferior temporal cortex subserve representational stability (Charest et al., 2014). These regions have, in fact, been implicated in computing invariant representations (Andrews & Ewbank, 2004; Baylis & Driver, 2001; Booth & Rolls, 1998; Rollenhagen & Olson, 2000). Our data, however, expands these findings, by suggesting that within-category stability is implemented, in part, within category-specific networks. Importantly, it has been shown that information flows in a long range fashion between nodes in a category-specific network, constraining and enriching local representations (Lee et al., 2019; Ruttorf, Kristensen, Schad, & Almeida, 2019). It may crucially be these global/distal influences on local representations that happen within category-specific networks (Lee et al., 2019; Ruttorf et al., 2019) that are at the basis of the kind of representational stability demonstrated here, and the kind of stability that allows for object individuation.

Moreover, our data also demonstrates that visuospatial processing of objects (e.g., in the right Angular gyrus and right superior parietal cortex; e.g., Chen et al. (2012), Peelle et al. (2009), Sabsevitz et al. (2005), Seghier (2013)) may be of special importance in conceptual representation and stability. Interestingly, it has been shown that patients with right parietal and fronto-parietal lesions leading to neglect can present with object-centered and allocentric

neglect syndromes (Chechlacz, Rotshtein & Humphreys (2012), Driver, Baylis, Goodrich, and Rafal (1994), Tipper and Behrmann (1996)) suggesting perhaps that the right angular gyrus contributes to 3D, viewer-independent representations of objects (for the importance of 3D viewer-independent representations on the perception of a particular class of objects – that of faces – see Almeida et al. (2020)).

Importantly, we also demonstrate that regions showing high representational stability hold representations that are dependent on information that is relatively detached from veridical and lower-level visual information. In fact, later layers of CNN, coding for more complex and abstract visual features, had been shown to be related with regions typically perceived as dedicated to more conceptual object processing (Cichy et al., 2016; Horikawa & Kamitani, 2017a; Khaligh-Razavi & Kriegeskorte, 2014). These results are consistent with previous findings showing that perceptual information is represented within temporal cortex (Mur et al., 2013; Peelen & Caramazza, 2012), whereas pixelwise information is represented in primary visual cortex (Bracci, Daniels, & Op de Beeck, 2017; Peelen & Caramazza, 2012). Note that CNN later layers and behavioral models on perceived similarity measure different things, but are, at least partially, both indexing higher-level representations specially when compared to earlier layers and models based on veridical similarity. Thus, our data suggests then that representational stability, as a central aspect of object individuation, requires a level of abstractness from sensorial information.

It is important to note that in this study we have tested only a particular category – that of tools – and thus did not extend our study to within-category stability for items belonging to other categories. Nevertheless, we predict that the regions showing stability for items from a particular category will include some of the nodes that are recruited for the processing of the target category, and specifically those that code for high-level abstract visual (or other sensorial) information. For instance, it putatively follows from these data that our ability to recognize our friends' faces or bodies at different times (and perspectives) will be, in part, dependent on areas such as the fusiform face area or the fusiform body area. This may be important to explore what types of information they represent and when they represent it.

Moreover, we only tested one particular (high-level) task – that of discriminating between a tool and a chimera. Whether the locus of representational stability is dependent on the task at hand is yet to be determined. In fact, it may well be that while the constitutive aspects of a concept (Mahon & Caramazza, 2008) are impervious to the task at hand, other kinds of information may be called upon (and show representational stability) under different tasks. Nevertheless, we set out to explore representational stability of object representations in general, and as such a high-level, object related task was perhaps the most suited for this task.

Recently, the use of CNNs as a proxy to object recognition has been challenged, as CNNs may not process visual stimuli in the same way as humans do. Particularly, it has been shown

that CNNs are more sensitive to noise, especially when dealing with outliers of unrecognizable images (Zhang, Liu, & Suen, 2020), and rely more on more local information (e.g., object texture) than global information (e.g., object shape; Baker, Lu, Erlikhman, and Kellman (2018), Geirhos et al. (2018)). This could perhaps be problematic as we try to compare the outputs of different layers (including lower-level layers that may be less similar to human visual processing than previously thought) against neuronal processing. Nevertheless, we show a consistency between these computational models and behavioral models, suggesting that these limitation in CNNs, albeit extremely important, may not be enough to prevent our conclusions here.

In sum, our study shows that representations are stable across time, and that within-category stability, potentially as a proxy of object individuation, is dependent nodes involved in the processing of the particular category of the presented object – in the case of tools, as tested here, regions within ventral temporal cortex. Moreover, the representations in high stability regions are dependent not on veridical visual or lower level visual (sensorial) information, but rather on high-level abstract visual properties – that is, the stability of object representations is achieved and consolidated through abstraction of visual (sensorial) information.

Materials and methods

The conditions of our ethics approval do not permit public archiving of anonymized study data. Readers seeking access to the data should contact the corresponding author through email. Full access to the data will be granted on request without conditions. Stimuli and code are available at https://osf.io/yx7rn/. No part of the study procedures and analysis was pre-registered prior to the research being conducted. We report how we determined our sample size, all data exclusions (if any), all inclusion/exclusion criteria, whether inclusion/exclusion criteria were established prior to data analysis, all manipulations, and all measures in the study.

Participants

Twenty healthy subjects (6 males and 14 females, ages from 19 to 43 years, mean = 22.1, SD = 5.4) participated in this study following previous studies (e.g., Almeida et al, 2013). All subjects were right-handed with normal or corrected-to-normal vision, and had no histories of neurological and psychiatric disorders. Sample size was determined given previous studies from the authors and from the extant literature. This study was approved by the ethical committee of the Faculty of Psychology and Educational Sciences, University of Coimbra, and written informed consent was obtained from all subjects.

Experimental design

All participants completed two fMRI sessions (plus another not analyzed herein) each separated by about a week (mean = 7.1 ± 0.6 days). During the fMRI sessions, participants performed an object recognition task composed of gray-scaled images of 80 tools (see Supplementary Figure S1for a list of the objects used) and 16 chimeras. In the object perception task, all stimuli were presented to participants using Psychtoolbox-3 (http://psychtoolbox.org). Participants were asked to discriminate whether each image is a tool or a chimera by pressing a button to ensure that they were focusing on the task. Each image was presented for 2s, separated by 4 seconds of fixation. Each participant completed 5 experimental runs per session. Across runs (odd and even) different exemplars of tools were used for each basic-level item. All stimuli were 400 × 400 pixels in size (~10° of visual angle; see Supplementary Figure S2 for examples of tools and chimeras used) and presented on a gray background using an Avotec projector (Stuart, FL, USA) under 60 Hz refresh rate.

Data acquisition and image processing

All MRI data were obtained from a Siemens 3.0T Tim Trio scanner (Berlin, Germany) with a 12-channel head coil at the Portuguese Brain Imaging Network. High-resolution structural T1-weighted data were obtained using a MPRAGE (magnetization prepared rapid gradient echo) sequence using the following parameters: MPRAGE (magnetization prepared

rapid gradient echo) sequence, a 256 × 256 acquisition matrix, a 230 mm field-of-view, a voxel size of 0.9 × 0.9 × 0.9 mm^3, a repetition time (TR) of 1900 ms, and an echo time (TE) of 2.32 ms. Functional magnetic resonance imaging (fMRI) data were acquired axially using T2*-weighted single-shot echo-planar imaging (EPI) sequence using the following parameters: a 96 × 96 acquisition matrix, a 220 mm field-of-view, 40 (interleaved) slices, a voxel size of 2.3 × 2.3 × 2.3 mm^3, a 2000 ms TR, a 22 ms TE, a flip angle of 90° and a gap of 0.7 mm.

Preprocessing of fMRI data was conducted using statistical parametric mapping (SPM12, http://www.fil.ion.ucl.ac.uk/spm/) (Friston et al., 1995). All functional data underwent standard preprocessing steps, including slice timing correction, head motion correction by realigning all consecutive volumes to the first image of each session, and co-registration of T1-weighted image to the first functional data using the linear registration algorithm. Co-registered T1-image was used to spatially normalize functional data into MNI template space using nonlinear transformation in SPM12. The functional data were interpolated to 2.0 × 2.0 × 2.0 mm^3 voxels. No spatial smoothing was conducted to avoid spill-over effects between voxels (Haxby et al., 2001; Haynes & Rees, 2006; Todd, Nystrom, & Cohen, 2013).

Regions of Interest (ROI)

For ROI-specific object representations, we parcellated the cerebral brain based on the human Brainnetome Atlas (http://atlas.brainnetome.org) (Fan et al., 2016). We defined 246 cerebral nodes in the individual structural space. To do this, we co-registered T1-weighted image to EPI using a linear registration algorithm between T1-weighted image and the fist EPI image in the first session on each day. The human Brainnetome Atlas in the MNI template space was transformed in to the individual T1-weighted image by applying the inverse nonlinear transformation using the DARTEL toolbox in SPM12 (Ashburner, 2007). The label map in the individual T1 space was spatially normalized to functional EPI space into MNI template space using non-linear transformation in SPM 12.

Representational similarity analysis

Representational similarity analysis (RSA, Kriegeskorte, Mur, Ruff, et al., 2008) was carried out using a condition-rich event-related fMRI experiment (Kriegeskorte, Mur, & Bandettini, 2008), where each of the 80 tools is treated as an experimental condition. We concatenated five continuous runs within each session and conducted a general linear model (GLM) analysis of fMRI data. The GLM ($Y = X\beta + \varepsilon$) was modeled by the weighted sum of a set of regressors for each of the 80 tools and the estimate of coefficient (β) that reflects voxel weights for brain activity for each tool. We converted the β estimates to t-values. The t-values were used to make neural activity patterns for 246 cortical regions. In each ROI, we calculated dissimilarity between a pair of neural patterns for each of the 80 tools using correlation distance

(i.e., 1–Pearson's *r*). Representational dissimilarity matrices (RDMs) were based on the dissimilarity of all pairs of neural patterns for 80 tools. To construct a general representational geometry for tool representations across participants, we averaged the RDMs of all participants per session. The group-average RDM was separately organized for each region and scanning day.

Convolutional neural network

It has been shown that hierarchical visual features of images used in fMRI experiments can be derived from a pre-trained convolutional neural network (CNN) (Horikawa & Kamitani, 2017a, 2017b). To carry out feature extraction from each image of the 80 tools, we utilized the MatConvNet toolbox (http://www.vlfeat.org/matconvnet/) (Vedaldi & Lenc, 2015). Specifically we extracted image features from each tool using AlexNet (Krizhevsky et al., 2012) pretrained on over a million images in ImageNet as the CNN model. The pre-trained CNN model, which can classify images into 1000 object categories, consisted of the five convolutional layers (CNN1–5) and the three fully connected layers for object classification. Each convolutional layer underwent typical CNN building, including linear filtering, non-linear transformation, max-pooling, and normalization. Fully connected layers (CNN6-7) were thresholded with a ratified linear unit (ReLU) and the last fully connected layer (CNN8) was fed with a softmax function. We obtained a vector of those units' outputs for each image and calculated the dissimilarity of all pairs of vectors for the 80 tools using correlation distance (i.e., 1−Pearson's r).

Human object similarity judgement

Sixty healthy, right-handed participants (19 males and 41 females, ages from 19 to 32 years, mean = 21.2, SD = 2.7) conducted an object-similarity judgment task composed of 80 tool words. Participants were asked to divide all stimuli into piles. When two tools are in the same pile similarity between a pair of tools is assigned to 1, and when they are not similarity is assigned to 0. Then we calculated the dissimilarity matrix by summing the assigned values per participant in each cell and subtracting the sum to the total number of participants.

Actual visual similarity analysis

For actual visual similarity analysis, we computed dissimilarity between the tool images using MATLAB 2-D correlation coefficient. Individual-specific tool images were compared, pixel by pixel, across sessions. Then, we constructed dissimilarity matrices for two fMRI sessions by applying correlation distance (i.e., 1−Pearson's r) and averaged them account for more robust pixelwise similarity.

Statistical comparison of representational geometries

We conducted permutations tests to generate a null distribution of correlation coefficients. For example, when two RDMs (e.g., neural RDM, CNN8 RDM) were given, the actual correlation was computed using Pearson correlation. And then, the stimulus labels of only one RDM were randomized (another RDM is fixed) and the correlation coefficient was calculated between the fixed RDM and the relabeled RDM. This step was repeated 10,000 times. If the actual correlation was within the top 5% of the null distributions of correlations with considering Bonferroni-corrected $p < 0.05$ for 246 regions, the null hypothesis that all stimuli had consistent patterns was rejected.

Authorship contribution statement

Dongha Lee: Conceptualization, Data curation, Data acquisition, Formal analysis, Writing - original draft, Writing - review & editing.

Jorge Almeida: Conceptualization, Data curation, Formal analysis, Funding acquisition, Supervision, Writing - original draft, Writing - review & editing.

Acknowledgements

This research was supported by a Foundation for Science and Technology of Portugal and Programa COMPETE grant (PTDC/MHC-PCN/0522/2014) and an European Research Council Starting Grant ("ContentMAP" - 802553) to JA, and by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (2018R1A6A3A03010805) to DL.

References

Almeida, J., Fintzi, A. R., & Mahon, B. Z. (2013). Tool manipulation knowledge is retrieved by way of the ventral visual object processing pathway. *Cortex, 49*(9), 2334-2344. doi: 10.1016/j.cortex.2013.05.004

Almeida, J., Freixo, A., Tabuas-Pereira, M., Herald, S. B., Valerio, D., Schu, G., . . . Santana, I. (2020). Face-Specific Perceptual Distortions Reveal A View- and Orientation-Independent Face Template. *Curr Biol*. doi: 10.1016/j.cub.2020.07.067

Almeida, J., Mahon, B. Z., Zapater-Raberov, V., Dziuba, A., Cabaco, T., Marques, J. F., & Caramazza, A. (2014). Grasping with the eyes: the role of elongation in visual recognition of manipulable objects. *Cogn Affect Behav Neurosci, 14*(1), 319-335. doi: 10.3758/s13415-013-0208-0

Andrews, T. J., & Ewbank, M. P. (2004). Distinct representations for facial identity and changeable aspects of faces in the human temporal lobe. *Neuroimage, 23*(3), 905-913. doi: 10.1016/j.neuroimage.2004.07.060

Ashburner, J. (2007). A fast diffeomorphic image registration algorithm. *Neuroimage, 38*(1), 95-113. doi: 10.1016/j.neuroimage.2007.07.007

Baker, N., Lu, H., Erlikhman, G., & Kellman, P. J. (2018). Deep convolutional networks do not classify based on global object shape. *PLoS Comput Biol, 14*(12), e1006613. doi: 10.1371/journal.pcbi.1006613

Baylis, G. C., & Driver, J. (2001). Shape-coding in IT cells generalizes over contrast and mirror reversal, but not figure-ground reversal. *Nat Neurosci, 4*(9), 937-942. doi: 10.1038/nn0901-937

Binkofski, F., & Buccino, G. (2006). The role of ventral premotor cortex in action execution and action understanding. *J Physiol Paris, 99*(4-6), 396-405. doi: 10.1016/j.jphysparis.2006.03.005

Booth, M. C., & Rolls, E. T. (1998). View-invariant representations of familiar objects by neurons in the inferior temporal visual cortex. *Cereb Cortex, 8*(6), 510-523.

Bracci, S., Daniels, N., & Op de Beeck, H. (2017). Task Context Overrules Object- and Category-Related Representational Content in the Human Parietal Cortex. *Cereb Cortex, 27*(1), 310-321. doi: 10.1093/cercor/bhw419

Chao, L. L., Haxby, J. V., & Martin, A. (1999). Attribute-based neural substrates in temporal cortex for perceiving and knowing about objects. *Nat Neurosci, 2*(10), 913-919. doi: 10.1038/13217

Chao, L. L., & Martin, A. (2000). Representation of manipulable man-made objects in the dorsal stream. *Neuroimage, 12*(4), 478-484. doi: 10.1006/nimg.2000.0635

Charest, I., Kievit, R. A., Schmitz, T. W., Deca, D., & Kriegeskorte, N. (2014). Unique semantic space in the brain of each beholder predicts perceived similarity. *Proc Natl Acad Sci U S A, 111*(40), 14565-14570. doi: 10.1073/pnas.1402594111

Chen, Q., Weidner, R., Vossel, S., Weiss, P. H., & Fink, G. R. (2012). Neural mechanisms of attentional reorienting in three-dimensional space. *J Neurosci, 32*(39), 13352-13362. doi: 10.1523/JNEUROSCI.1772-12.2012

Chechlacz, M., Rotshtein, P., & Humphreys, G. W. (2012). Neuroanatomical Dissections of Unilateral Visual Neglect Symptoms: ALE Meta-Analysis of Lesion-Symptom Mapping. *Frontiers in human neuroscience, 6*, 230. https://doi.org/10.3389/fnhum.2012.00230

Cichy, R. M., Khosla, A., Pantazis, D., Torralba, A., & Oliva, A. (2016). Comparison of deep neural networks to spatio-temporal cortical dynamics of human visual object recognition reveals hierarchical correspondence. *Sci Rep, 6*, 27755. doi: 10.1038/srep27755

Cree, G. S., & McRae, K. (2003). Analyzing the factors underlying the structure and computation of the meaning of chipmunk, cherry, chisel, cheese, and cello (and many other such concrete nouns). *J Exp Psychol Gen, 132*(2), 163-201. doi: 10.1037/0096-3445.132.2.163

DiCarlo, J. J., & Cox, D. D. (2007). Untangling invariant object recognition. *Trends Cogn Sci, 11*(8), 333-341. doi: 10.1016/j.tics.2007.06.010

Driver, J., Baylis, G. C., Goodrich, S. J., & Rafal, R. D. (1994). Axis-based neglect of visual shapes. *Neuropsychologia, 32*(11), 1353-1365. doi: 10.1016/0028-3932(94)00068-9

Fan, L., Li, H., Zhuo, J., Zhang, Y., Wang, J., Chen, L., . . . Jiang, T. (2016). The Human Brainnetome Atlas: A New Brain Atlas Based on Connectional Architecture. *Cereb Cortex, 26*(8), 3508-3526. doi: 10.1093/cercor/bhw157

Farrer, C., Frey, S. H., Van Horn, J. D., Tunik, E., Turk, D., Inati, S., & Grafton, S. T. (2008). The angular gyrus computes action awareness representations. *Cereb Cortex, 18*(2), 254-261. doi: 10.1093/cercor/bhm050

Friston, K., Holmes, A., Worsley, K., Poline, J., Frith, C., & Frackowiak, R. (1995). Statistical parametric maps in functional imaging: a general linear approach. *Human Brain Mapping, 2*, 189-210.

Garcea, F. E., Kristensen, S., Almeida, J., & Mahon, B. Z. (2016). Resilience to the contralateral visual field bias as a window into object representations. *Cortex, 81*, 14-23. doi: 10.1016/j.cortex.2016.04.006

Geirhos, R., Rubisch, P., Michaelis, C., Bethge, M., Wichmann, F. A., & Brendel, W. (2018). ImageNet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness. *arXiv preprint arXiv:1811.12231*.

Gerlach, C., & Marques, J. F. (2014). Visual complexity exerts opposing effects on object categorization and identification. *Visual Cognition, 22*(6), 770-788. doi: 10.1080/13506285.2014.915908

Grill-Spector, K., Kushnir, T., Edelman, S., Avidan, G., Itzchak, Y., & Malach, R. (1999). Differential processing of objects under various viewing conditions in the human lateral occipital complex. *Neuron, 24*(1), 187-203.

Haxby, J. V., Gobbini, M. I., Furey, M. L., Ishai, A., Schouten, J. L., & Pietrini, P. (2001). Distributed and overlapping representations of faces and objects in ventral temporal cortex. *Science, 293*(5539), 2425-2430. doi: 10.1126/science.1063736

Haynes, J. D., & Rees, G. (2006). Decoding mental states from brain activity in humans. *Nat Rev Neurosci, 7*(7), 523-534. doi: 10.1038/nrn1931

Horikawa, T., & Kamitani, Y. (2017a). Generic decoding of seen and imagined objects using hierarchical visual features. *Nat Commun, 8*, 15037. doi: 10.1038/ncomms15037

Horikawa, T., & Kamitani, Y. (2017b). Hierarchical Neural Representation of Dreamed Objects Revealed by Brain Decoding with Deep Neural Network Features. *Front Comput Neurosci, 11*, 4. doi: 10.3389/fncom.2017.00004

Khaligh-Razavi, S. M., & Kriegeskorte, N. (2014). Deep supervised, but not unsupervised, models may explain IT cortical representation. *PLoS Comput Biol, 10*(11), e1003915. doi: 10.1371/journal.pcbi.1003915

Konen, C. S., & Kastner, S. (2008). Two hierarchically organized neural systems for object information in human visual cortex. *Nat Neurosci, 11*(2), 224-231. doi: 10.1038/nn2036

Kriegeskorte, N., Mur, M., & Bandettini, P. (2008). Representational similarity analysis - connecting the branches of systems neuroscience. *Front Syst Neurosci, 2*, 4. doi: 10.3389/neuro.06.004.2008

Kriegeskorte, N., Mur, M., Ruff, D. A., Kiani, R., Bodurka, J., Esteky, H., . . . Bandettini, P. A. (2008). Matching categorical object representations in inferior temporal cortex of man and monkey. *Neuron, 60*(6), 1126-1141. doi: 10.1016/j.neuron.2008.10.043

Kristensen, S., Garcea, F. E., Mahon, B. Z., & Almeida, J. (2016). Temporal Frequency Tuning Reveals Interactions between the Dorsal and Ventral Visual Streams. *J Cogn Neurosci, 28*(9), 1295-1302. doi: 10.1162/jocn_a_00969

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 1097-1105.

Lee, D., Mahon, B. Z., & Almeida, J. (2019). Action at a distance on object-related ventral temporal representations. *Cortex, 117*, 157-167. doi: 10.1016/j.cortex.2019.02.018

Mahon, B. Z., & Caramazza, A. (2008). A critical look at the embodied cognition hypothesis and a new proposal for grounding conceptual content. *J Physiol Paris, 102*(1-3), 59-70. doi: 10.1016/j.jphysparis.2008.03.004

Mahon, B. Z., Kumar, N., & Almeida, J. (2013). Spatial frequency tuning reveals interactions between the dorsal and ventral visual systems. *J Cogn Neurosci, 25*(6), 862-871. doi: 10.1162/jocn_a_00370

Mahon, B. Z., Milleville, S. C., Negri, G. A., Rumiati, R. I., Caramazza, A., & Martin, A. (2007). Action-related properties shape object representations in the ventral stream. *Neuron, 55*(3), 507-520. doi: 10.1016/j.neuron.2007.07.011

Marques, J. F., Raposo, A., & Almeida, J. (2013). Structural processing and category-specific deficits. *Cortex, 49*(1), 266-275. doi: 10.1016/j.cortex.2011.10.006

Mur, M., Meys, M., Bodurka, J., Goebel, R., Bandettini, P. A., & Kriegeskorte, N. (2013). Human Object-Similarity Judgments Reflect and Transcend the Primate-IT Object Representation. *Front Psychol, 4*, 128. doi: 10.3389/fpsyg.2013.00128

Noppeney, U., Price, C. J., Penny, W. D., & Friston, K. J. (2006). Two distinct neural mechanisms for category-selective responses. *Cereb Cortex, 16*(3), 437-445. doi: 10.1093/cercor/bhi123

Peelen, M. V., & Caramazza, A. (2012). Conceptual object representations in human anterior temporal cortex. *J Neurosci, 32*(45), 15728-15736. doi: 10.1523/JNEUROSCI.1953-12.2012

Peelle, J. E., Troiani, V., & Grossman, M. (2009). Interaction between process and content in semantic memory: an fMRI study of noun feature knowledge. *Neuropsychologia, 47*(4), 995-1003. doi: 10.1016/j.neuropsychologia.2008.10.027

Pourtois, G., Schwartz, S., Spiridon, M., Martuzzi, R., & Vuilleumier, P. (2009). Object representations for multiple visual categories overlap in lateral occipital and medial fusiform cortex. *Cereb Cortex, 19*(8), 1806-1819. doi: 10.1093/cercor/bhn210

Riesenhuber, M., & Poggio, T. (2002). Neural mechanisms of object recognition. *Curr Opin Neurobiol, 12*(2), 162-168. doi: 10.1016/s0959-4388(02)00304-5

Rollenhagen, J. E., & Olson, C. R. (2000). Mirror-image confusion in single neurons of the macaque inferotemporal cortex. *Science, 287*(5457), 1506-1508.

Ruttorf, M., Kristensen, S., Schad, L. R., & Almeida, J. (2019). Transcranial Direct Current Stimulation Alters Functional Network Structure in Humans: A Graph Theoretical Analysis. *IEEE Trans Med Imaging, 38*(12), 2829-2837. doi: 10.1109/TMI.2019.2915206

Sabsevitz, D. S., Medler, D. A., Seidenberg, M., & Binder, J. R. (2005). Modulation of the semantic system by word imageability. *Neuroimage, 27*(1), 188-200. doi: 10.1016/j.neuroimage.2005.04.012

Sakuraba, S., Sakai, S., Yamanaka, M., Yokosawa, K., & Hirayama, K. (2012). Does the human dorsal stream really process a category for tools? *J Neurosci, 32*(11), 3949-3953. doi: 10.1523/JNEUROSCI.3973-11.2012

Seghier, M. L. (2013). The angular gyrus: multiple functions and multiple subdivisions. *Neuroscientist, 19*(1), 43-61. doi: 10.1177/1073858412440596

Serre, T., Oliva, A., & Poggio, T. (2007). A feedforward architecture accounts for rapid categorization. *Proc Natl Acad Sci U S A, 104*(15), 6424-6429. doi: 10.1073/pnas.0700622104

Tipper, S. P., & Behrmann, M. (1996). Object-centered not scene-based visual neglect. *J Exp Psychol Hum Percept Perform, 22*(5), 1261-1278. doi: 10.1037//0096-1523.22.5.1261

Todd, M. T., Nystrom, L. E., & Cohen, J. D. (2013). Confounds in multivariate pattern analysis: Theory and rule representation case study. *Neuroimage, 77*, 157-165. doi: 10.1016/j.neuroimage.2013.03.039

Tyler, L. K., & Moss, H. E. (2001). Towards a distributed account of conceptual knowledge. *Trends Cogn Sci, 5*(6), 244-252. doi: 10.1016/s1364-6613(00)01651-x

Vedaldi, A., & Lenc, K. (2015). MatConvNet: Convolutional neural networks for MATLAB. *Proceedings of the 23rd ACM International Conference on Multimedia (ACM, 2015)*, 689–692.

Warrington, E. K., & McCarthy, R. A. (1987). Categories of knowledge. Further fractionations and an attempted integration. *Brain, 110 (Pt 5)*, 1273-1296. doi: 10.1093/brain/110.5.1273

Zhang, X. Y., Liu, C. L., & Suen, C. Y. (2020). Towards Robust Pattern Recognition: A Review. *Proceedings of the IEEE, 108*(6), 894-922.

Table 2. Similarity of neural RDMs in high within-category representational areas with CNN, perceived and veridical RDMs.

ROIs	Anatomical and modified Cytoarchitectonic descriptions from Brainnetome Atlas	CNN1	CNN2	CNN3	CNN4	CNN5	CNN6	CNN7	CNN8	Perceived	Veridical
Early visual cortex	L.OPC (occipital polar cortex)	-0.3361	-0.1043	0.1670	0.0631	-0.0325	-0.0374	-0.1556	-0.1576	0.0238	0.4558
	R.OPC	-0.2515	-0.0686	0.1784	0.0828	-0.0156	-0.0111	-0.1283	-0.1508	0.0054	0.4506
	R.mOccG (middle occipital gyrus)	-0.1952	-0.0194	0.0855	0.0367	-0.0304	0.1311	0.0523	0.0510	0.0374	0.0974
Posterior middle temporal gyrus	L.V5/MT+ (area V5/MT+)	-0.0249	0.1404	0.1019	0.1086	0.1009	0.3155	0.3025	0.2815	0.0712	-0.1236
	R.V5/MT+	-0.0058	0.1354	0.0770	0.0712	0.0664	0.2892	0.2807	0.2776	0.0560	-0.1761
Fusiform gyrus	L.A37lv (lateroventral area 37)	0.0928	0.1937	0.0808	0.1063	0.1232	0.2732	0.2880	0.2847	0.0790	-0.1908
Superior parietal lobule	R.A7ip (intraparietal area 7, hIP3)	0.1342	0.2140	0.0569	0.1015	0.1316	0.2296	0.2532	0.2409	0.0752	-0.2843
Inferior parietal lobule	R.A39c (caudal area 39, PGp)	0.0538	0.0955	0.0199	0.0426	0.0347	0.1828	0.1702	0.1564	0.0547	-0.2004
	R.A39rd	0.2002	0.2636	0.0775	0.1332	0.1722	0.2730	0.2925	0.2661	0.0739	-0.2840
Ventral premotor cortex	R.A6cvl (caudal ventrolateral area 6)	0.1702	0.2313	0.0615	0.1022	0.1264	0.2192	0.2467	0.2369	0.0734	-0.2404
Dorsal premotor cortex	L.IFJ (inferior frontal junction)	0.1160	0.1945	0.0594	0.0998	0.1249	0.2089	0.2299	0.2077	0.0727	-0.1875
	R.IFJ	0.1014	0.2049	0.0622	0.0944	0.1188	0.2023	0.2258	0.2281	0.0759	-0.2293